EvoFIT: A holistic, evolutionary facial imaging technique for creating composites

Frowd, Charlie, Hancock, Peter J. B. and Carson, Derek

Frowd, Charlie ORCID: 0000-0002-5082-1259, Hancock, Peter J. B. and Carson, Derek (2004) EvoFIT: A holistic, evolutionary facial imaging technique for creating composites. ACM Transactions on Applied Perception (TAP), 1 (1). pp. 19-39. ISSN 1544-3558

EvoFIT: A Holistic, Evolutionary Facial Imaging Technique for Creating Composites

CHARLIE D. FROWD[1], PETER J.B. HANCOCK[2] and DEREK CARSON[3]

ABSTRACT

EvoFIT, a computerized facial composite system is being developed as an alternative to current systems. EvoFIT Faces are initially presented to a witness with random characteristics, but through a process of selection and breeding, a composite is 'evolved'. Comparing composites constructed with E-FIT, a current system, a naming rate of 10% was found for EvoFIT and 17% for E-FIT. Analysis revealed that target age was limiting factor for EvoFIT and a second study with age appropriate targets visible during composite construction produced a naming rate similar to E-FIT. Two more-realistic studies were conducted that involved young target faces and 2 current systems (E-FIT and PROfit). Composites from both of these experiments were poorly named but a significant benefit emerged for EvoFIT.

Keywords: Design, Experimentation, Human factors.

To appear in ACM Transactions on Applied Perception

---

[1] Department of Psychology, University of Stirling, FK9 4LA. cdf1@stir.ac.uk

[2] Department of Psychology, University of Stirling, FK9 4LA. pjbh1@stir.ac.uk

[3] Division of Psychology, University of Abertay, DD1 1HG. d.carson@mail1.tay.ac.uk

Contents

1. INTRODUCTION

Suppose you are witness to a serious crime. The police will want to know what the perpetrator looks like and may ask you to try to produce an image of his or her face. This will most likely be with the aid of a computer based facial composite program. These contain numerous photographs of bits of faces (eyes, noses etc), which can be pasted together to make a whole face. Such systems can be very flexible and under ideal conditions, namely with the target or a photograph present during construction, are capable of producing very convincing images. However, when working from memory, as would be the case with a crime, performance is far worse, with the composites being recognised only about 20% of the time at best [Bruce et al. 2002; Davies et al. 2000].

Part of the problem may well be fundamental to the way the composite systems work. Witnesses are asked to concentrate on individual facial features, when there is much psychological evidence to show that we perceive faces as wholes. Change the nose in a face and an observer may well be able to tell you that something has changed but not necessarily what. It is therefore rather difficult to say what needs to be changed to make a given face more like a particular target.

This work describes a full implementation of the novel face generation system described in pilot form in Hancock [2000]. This takes a radically different approach to generating the face image. Rather than assembling individual features, the system combines "eigenfaces", each of which may affect the appearance of the whole face. Rather than asking a witness what needs to be changed, the system presents a number of possible faces and asks the witness to select those that look most like the target. An underlying evolutionary algorithm then generates a new set of faces from those selected. This process is iterated through a number of generations, gradually moving closer to the target face.

1.1. Eigenfaces

Over the last decade, principal components analysis (PCA) has been used extensively for the analysis and recognition of faces. PCA is a well-established statistical method for simplifying large datasets. Applied to face images, it produces ghostly "eigenfaces" which capture the major modes of variation within the image set. Sirovich and Kirby [1987] were the first to demonstrate that the technique works well for faces. They started with monochrome photographs of 115 full-face Caucasian males. Simple normalization was performed that aligned the head in the vertical plane, the eyes in the horizontal plane and resized the image to make the width of the head the same in each photograph. Because the early components capture most of the variance, they found that the faces could be well represented (to within 4% average error) using only the first 50.

Because Sirovich and Kirby aligned faces only rather crudely, the features within each face were in different locations. The PCA was performed on the image pixels, so the eigenfaces faithfully coded this variation. If these eigenfaces are recombined in proportions other than those found in the original images, they will tend to give faces that have blurred features. The answer is more sophisticated alignment of features [Brunelli and Poggio 1993; Craw and Cameron 1991; Troje and Vetter 1996]. For example, Craw and Cameron located co-ordinate or "control points" around the major facial features (eyes, eyebrows, nose, mouth) and the outline of the head, including the ears, chin and jaw. The average position of each control point was computed across the image set and the image was triangulated to produce an image mesh. Each database image was then morphed to the average face shape before performing PCA. This is achieved by distorting the areas of the image defined by triangles (a bilinear interpolation) such that equivalent triangles in each face were made to have the same shape. The resultant images are referred to as *shape-free*. In their study, they demonstrate that faces not part of this image database can be constructed with "almost identical" accuracy using a linear combination of eigenfaces.

Hancock et al. [1996] argue that the control point information can itself form part of a PCA that models the relational aspects of the face (e.g. the distances between facial features). They refer to this as a *shape* model and the resulting eigenvectors are termed *eigenshapes*. The second model, concerning the *shape-free* image intensities, is referred to as the *texture* model. The term *texture* is used in a restricted sense, referring to the information in the image that remains after the face has been made *shape-free*. Examples of faces generated from the shape model can be seen in Figure 1, displayed with the average texture of the image corpus. Also displayed are examples of faces generated from the texture model. Cootes et al. [2000] combine the shape and texture models into a single appearance model. We keep them separate, as we have found that users like to work on the shape and texture of a face independently. By suitable adjustment of the model parameters (i.e. the proportions of each individual eigenface and eigenshape to be used) it is possible to produce novel faces from within the space defined by the original face images.

We therefore have a face model that can, potentially, produce faces to order. The problem is how to find an appropriate set of parameters. Giving the user a set of "knobs" to twiddle is not satisfactory for the reasons outlined above: it is too hard to work out what to change to make a face more like the target. This is partly because, although the principal components do show significant correlations with human perceptions of the variations within faces [Hancock et al. 1996; 1998; O'Toole et al. 1997], the correlations are weak and the components are not easily given labels (such as roundness of face).

(a) Facial Shape

(b) Facial Texture

Fig. 1. Examples of facial shapes and facial textures from the EvoFIT face model. These images were created in the initial generation and therefore have random characteristics. The hair was imported from the PROfit composite system (as normal). The facial shapes are shown with the average database texture, the facial textures all have the same facial shape (the average database shape).

The solution proposed is an evolutionary one. An initial set of faces is generated by using random numbers for the model parameters. The witness is then asked to select those that look most like the target. An Evolutionary Algorithm (EA) takes the selected parameter sets and produces a new set of faces, by recombination and mutation. The new faces will combine aspects of those previously chosen and is likely to include some that look more like the target. The process is repeated with, usually, gradual convergence towards the desired target. The difference in approach to conventional composite systems is highly significant: it taps into our rather good ability to recognise faces rather than our rather poor ability to describe them.

An evolutionary approach to composite generation has been described previously [Caldwell and Johnston 1991]. However, their system manipulates traditional composite features. The system codes which nose, eyes etc are used and recombination will produce faces that may use the eyes from one and the nose from another. Unfortunately there is no easy way to produce a similarity metric for the different features. Ideally, you would like a small change in the nose parameter to produce a small change in the nose, but without any kind of ordering of the nose images, any mutation will simply result in a different nose. Producing an ordering would require exhaustive evaluation of the similarity spaces for each type of feature. Even then, the space is bound to be multi-dimensional (nose length, width, shape etc), limiting the possible smoothness of the mapping. The PCA model does not suffer from this problem: a small change in any parameter will produce a correspondingly small image change.

Another weakness of current composite systems is the need for an operator to interpret what the witness is saying. Brace et al. [2000] found that the recognition of composites created with E-FIT was about 10% higher when their operator was also the “witness”, thereby eliminating the need to transfer information. Allowing witnesses to select better faces should sidestep this problem. Rakover and Cahlon [1996] presented pairs of composites sampled from the Identikit system and participants selected the more preferable. The system ultimately composed a composite based on the features in the faces that were selected most often. They found that when the target is present, their approach resulted in over 75% of the “correct” features being assembled in the composite. However, neither this approach, nor Caldwell and Johnstone’s, have been formally compared using composites constructed from memory. It is possible that presentation of too many faces will confuse the witness.

Hancock [2000] used a shape and texture PCA model from 20 Caucasian female faces as part of his prototype composite system. The approach was similar to Caldwell and Johnston [1991] in that a composite was created by interactive evolution. Here we extend this work and describe a full composite system (EvoFIT).

1.2. The EvoFIT System

The approach adopted employs the key features of Hancock [2000]: generation using a PCA shape and texture face model, the parallel presentation of faces and an EA for “breeding” new populations. Rather than individually rating each face, which took too long, users are asked to select a small number of faces (typically six) from a larger set (typically 18). They are also asked to nominate a best face, which is then given additional weighting within the EA and can also be used to help evaluate the system’s performance.

Construction of EvoFIT’s face model proceeds, as in Craw and Cameron [1991], with the manual location of facial feature boundaries using “control points” – 240 in total; an example can be seen in Figure 2.

A total of 72 monochrome photographs of young adult Caucasian males were used, these being extracted from an image corpus provided by the U.K. Home Office. Images are available in full-face pose with a neutral expression and under controlled lighting. As in Hancock [2000], two models were constructed with this image repository, one for facial shape and the other for facial texture. Generation of a novel face is carried out as above, by generating a random facial shape and using it to morph a random facial texture.

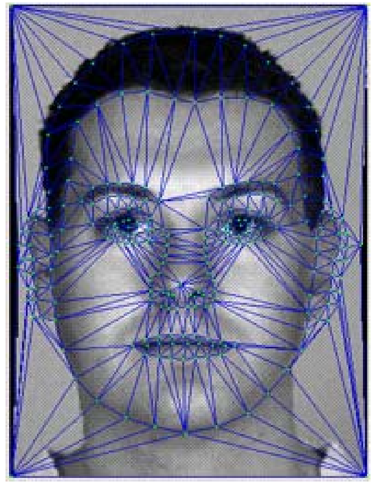

Fig. 2. Alignment of facial feature co-ordinate points. An image mesh is also shown, connecting the co-ordinate points as triangles. Distortion of the triangles (a bilinear interpolation, or *morph*) enables different facial shapes to be produced.

While the two-stage PCA model handles the general shape and the internal features of a face well, it does not cope with typical variations in hairstyle. Such variation includes situations where criminals deliberately change their appearance following a crime. There is no reliable association then between the appearance of a face and a hairstyle. Therefore, it is sensible to treat the hair as a free parameter independent of the shape and texture PCA models. In the pilot system this was handled by ensuring that all the hairstyles were very similar and short. EvoFIT has an interface to a standard computerised composite system such as PROfit that enables the importation of a hairstyle. Use of a composite method for the hair is consistent with the informal observation that people find it relatively easy to describe a person’s hair. The process involves first exporting a reference face (a face with average shape and texture) into PROfit. An appropriate hairstyle is applied to this prototype image and then imported back into EvoFIT. Appropriate facial textures are applied to this updated reference image (from the texture model) and the image morphed by control points defined by the shape model. To accommodate flexibility, the hair may be changed at any time. Figure 3 illustrates the effect of adding different hairstyles to a population face.

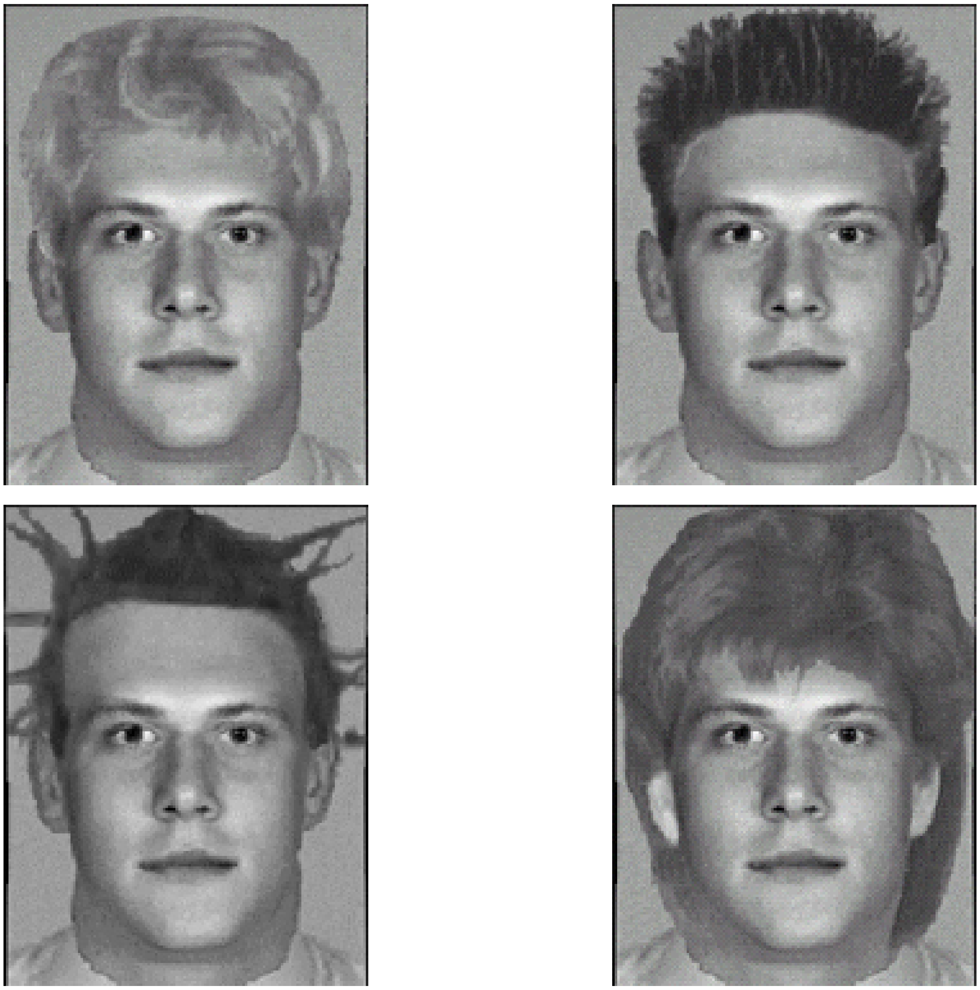

Fig. 3. Hairstyles imported from the PROfit composite system and applied to a population face.

Although an extensive range of hairstyles is available in PROfit, it is sometimes necessary to modify the chosen hair. In addition, it may also be necessary to add “adornments” such as an earring, an ear stud or a necklace. Provision was made therefore to allow modification of the “external facial features” in several standard photographic editing packages: Microsoft Paint, Microsoft PhotoEditor and Adobe Photoshop. In practice, a simple utility was designed that transfers a reference image into an editor and updates the population faces to reflect any changes made by an operator. The utility of the EvoFIT system was further expanded by allowing moles, scars, beauty marks and even adornments to be added to the “internal facial features”. These features are also added through one of the image editors, although internally these changes are maintained via an “overlay” mask. An example of the utility of this tool can be seen in Figure 4.

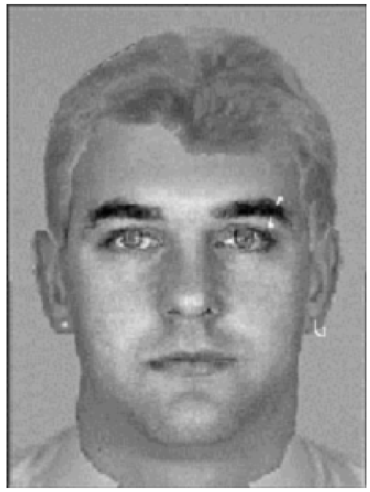

Fig. 4. An example of 'adornments' added to the internal and external facial features.

Setting system parameters

The underlying evolutionary algorithm works on the shape and texture PCA coefficients that define each face. These coefficients are held as real numbers (rather than the bit strings of traditional Genetic Algorithms). The initial values are obtained from a Gaussian random number generator, with standard deviation scaled to the variance of the component concerned (this ensures that plausible shapes and textures are generated).

There are a number of free parameters associated with the algorithm, such as mutation rate and population size. Ideally, good values for these would be established using repeated trials of the full system. However, this would be inordinately time-consuming and so a computer simulation approach was adopted (full details in Frowd [2001]). The simulation process replaced the normal selection of faces (by a witness) with an automatic one that chose faces with the smallest pixel mean squared error to a given target. While not a perfect match, tests found that estimates of composite quality (a rating of 'likeness' to the target face) did significantly correlate with the error measure ($r = -0.31$), suggesting that it may serve as a proxy to human choices. Simulations were run for 70 targets to achieve a good measure of average performance. Using this evaluation method allowed multiple runs, varying parameters. It was found:

- The system was insensitive to population size (the number of displayed faces) over the range 10-32. Specifically, it was found that a smaller population size run for more generations resulted in equivalent performance to larger population run for relatively a shorter time. Initially, the population size was set to 18 faces – the number of faces that could be comfortably viewed on a computer monitor.

- The optimal mutation rate was 0.1. This is the probability with which the Evolutionary Algorithm (EA) replaces each coefficient from the shape and texture model with a random value. A setting of 0.1 therefore replaces, on average, about 7 of the 71 coefficients. This is a relatively high mutation rate, needed to increase diversity of the rather small population.

- It was advantageous to use an "elitist" strategy. This approach, common in EAs, carries forward the best individual(s) from one generation to the next, preventing their loss through recombination and mutation. The best face chosen by the user therefore appears unchanged in the next selection of faces.

- It was also advantageous to double the influence of the "best face" relative to other population faces. In brief, each selected face was given an equal opportunity to become a parent (and therefore participate in the breeding process), except the best face, which enjoyed twice as many breeding opportunities.

- Faster evolution resulted when fewer faces were selected from each population. This was found to hold down to a minimum of 4 faces. In practice then, a witness should be careful in the number of faces selected.

Extensions to the basic system

Preliminary evaluation of the system revealed that sometimes a population face was generated with a good texture to a target but not a good shape (and vice versa). When this happened, a witness would have to grudgingly accept a poor quality representation. This was overcome by separating the shape and texture information onto different screens (referred to as different "palettes") for selection, called the Facial Shape Palette (FSP) and Facial Texture Palette (FTP) respectively. The procedure for constructing a composite then begins with the selection of facial shapes (on the FSP) and then facial textures (on the FTP). For the first generation of faces, faces on the FTP are presented with average shape since none has been chosen at this point, but thereafter, the shape of the best face selected in the previous generation is used. Similarly, images on the FSP are displayed with average texture in the first generation and then with the texture taken from the best face for later generations. Once some face shapes and textures have been chosen, combinations of them are displayed on a third screen to allow selection of the overall best face for that generation. If a user would prefer a different combination of shape and texture, this may be specified.

Preliminary work also revealed that it can be difficult to find enough faces sufficiently similar to the target from the eighteen presented. This is addressed by generating further sets of faces on demand.

Although we have argued that it can be difficult to specify what needs to be done to improve a face, it is sometimes the case that a witness may be able to. For example, they may wish to move the eyes closer together. EvoFIT has a Feature Shift Tool (FST) to enable this. It first selects those control points delineating a facial feature (e.g. the eyes). The control points can then be moved or resized as necessary. Each time, the utility carries out a morph on the face to reflect the change. In this way, the FST works very much like the computerised composite systems, repositioning and resizing features. The difference is that changes to the control points are followed by a best fit in the shape model and therefore manipulations are carried out in the holistic shape space. In practice, changes are made to the shape vector which is then projected back into the shape space (a best fit on all the control points). This serves to limit implausible relationships between features (as mentioned earlier). It also means that the changes are made in the "genetic code", so that they are propagated into succeeding generations (i.e. this 'improved' face is used for breeding). Figure 5 illustrates the effect of changing the horizontal spacing between the eyes using this utility. Note that the "best fit" approach identifies the closest face in the shape model and may result in small pixel movements in other areas of the composite (necessary to minimise the overall error).

Fig. 5. An example illustrating the effect of moving the eyes closer into register (by 8 pixels). The manipulated image (right) also illustrates that small movements can occur elsewhere in the face when performing a best fit in the shape model (e.g. smaller nostrils).

To summarize, the process of creating a composite with EvoFIT begins by selecting an appropriate hairstyle (e.g. from PROfit). The system generates random faces that are blended with the chosen hair. Four to six facial shapes and facial textures are selected along with an overall "best face". These choices are bred together by an Evolutionary Algorithm to produce another set of faces. The selection and breeding continues until an acceptable likeness is reached. Any of the population faces can be modified by manipulating the size and location of any feature (in the Feature Shift Tool). There is also the ability to combine the shape from one face and the texture from another.

2. EXPERIMENT 1

To gain a good measure of average performance, EvoFIT composites should be constructed from a large range of targets. To be realistic, these constructions should be carried out from memory and compared with performance from another current composite system.

2.1. Method

The most realistic method of construction is to allow participants to view a staged crime and then to create a composite. In this case, it would be necessary for the “assailant” to be unknown to the participants. To maintain an ecologically valid evaluation, the resulting composites would need to be shown to people who know the “assailant”. In addition, to obtain a measure of general performance, a large number of targets should be employed. For example, Kovera et al. [1997] created over 50 composites. The problem with staging crimes containing this number of targets is that it is rather time-consuming. As a compromise, it was decided to create composites of generally well-known people, such as famous actors and musicians. Although the use of such stimuli is not entirely ecologically valid, given that the targets are likely to be familiar, they can nevertheless be created from memory. Curiously, the effect of target familiarity (i.e. the amount of exposure to a target) may not be an issue anyway when constructing composites from memory, as Davies et al. [2000] discovered with both E-FIT and Photofit.

Although it may be not an issue for these systems, the effect of familiarity is unknown for EvoFIT. To this end, it was decided to that participants should receive a 1 minute exposure of their chosen target. This was aimed at counteracting target-exposure and retention-interval effects; that is, limiting differences in familiarity and the last time the famous person was seen. All participants would therefore begin the composite process with an equivalent target exposure and be less dependent on the last time they saw the famous face.

Even controlling for the duration of exposure and the retention interval, it was believed that the overall level of “distinctiveness” of a target could affect the quality of a composite. Distinctiveness is a measure of deviation from the average in a given set of faces. It is well established in face perception that distinctive faces are better recognized than more typical ones [e.g. Hancock et al. 1996; Shapiro and Penrod 1986; Valentine and Endo 1992]. For example, Light et al. [1979] report a significant increase in accuracy (an increase in hit rate and a decrease in false alarms) on a recognition task for distinctive faces compared with more average looking exemplars.

As the memory for distinctive faces is better, one would expect composites of distinctive faces to be better recognised. Curiously though, such an effect has yet to be demonstrated. For example, Green and Geiselman [1989] found composites made with the Identikit system were identified at chance level (in a simple 6 item photo line-up) when distinctive targets were used. However, this study featured the older line-based version of Identikit and performance might have been better had a more realistic representation been used. It was thought interesting then to investigate whether distinctiveness was a factor with EvoFIT; a system that does produces more realistic looking composites.

Now, although the effect of familiarity may not be a major issue for composite construction, as discussed above, familiarity is likely to be important during the recognition of a composite. It was decided therefore to keep the level of familiarity as constant as possible, but to manipulate the level of target distinctiveness in a set of famous face targets.

Another advantage of using famous faces as targets is that it fits in with current research. Recent work in this area includes Brace et al. [2000] and Davies and Oldman [1999]. Both of these studies constructed composites using E-FIT from memory and with the target present. Brace et al. [2000] found a recognition rate of 25% by presenting pairs of composites from both memory and target-present conditions; the target-present composites were made by the operator alone and the composites from memory were obtained in the normal way with a “describer” and an operator. Davies and Oldman [1999] found that individual composites were recognized 6% in the memory condition and 10% in the target present condition (the target was re-introduced for the witness to suggest changes). Using these studies as a guide, one would expect quite a wide a recognition rate: between 6% and 25%, although the upper limit is likely to be somewhat lower since Brace et al. [2000] presented multiple composites for recognition, a format known to elevate performance [Bennett 2000; Bruce et al. 2002].

An important aspect when designing a new system is, where possible, to demonstrate performance against another current system(s). The main composite systems used in the UK are PROfit and E-FIT. These are very similar systems in design and use. Both contain a large collection of photographed features for witnesses to select (e.g. hair, face shape, eyes, nose, etc.) In practice, witnesses first describe the features of a face (via a Cognitive Interview, see below) and then select features to match. Each of these chosen features may also be resized and positioned as required. Sometimes, only a general likeness is produced, given limitations in the feature database, and therefore a paint package is often used to enhance the quality (especially for older, more distinctive faces).

It was decided to perform a comparison between E-FIT and EvoFIT as experienced operators were available. These operators were authors of the current paper: first author (EvoFIT) and third author (E-FIT). To this end, composites of the same set of targets were constructed by both systems.

2.2. Obtaining the target faces

As mentioned above, one of the objectives of the study was to test performance with a relatively large number of targets. Ultimately, 30 were believed sufficient to gain a good measure of average performance. To ensure that the operators could not initially bias the construction of the composites, they were unaware of the identity of the targets.

Caucasian male targets were used, so as to fit the design of the current EvoFIT database. Thirty monochrome photographs of white famous individuals were assembled, including actors, sportsmen, singers and TV personalities well known to residents in the UK. Each person was depicted in (as far as possible) a full-face pose and a neutral expression. These photographs were rated for distinctiveness and familiarity (i.e. how well the person was known), which enabled the set to be divided into three different distinctiveness levels (low, medium and high) with equivalent familiarity.

2.3. Creating the Composites

A procedure used in the UK to train police operators to elicit information from a witness is based on a “cognitive approach” [FIC 1999]. This approach, used during a Cognitive Interview (CI), is designed to facilitate the recall of as much unbiased information as possible regarding a crime, largely through re-instating the context in which the event took place. Part of the CI involves eliciting a verbal description of the suspect, including the face. The verbal description typically involves a phase whereby a witness recalls (and then re-recalls) details of the event in his or her own time with the minimum of external cueing; referred to as “free-recall”. This is followed by a more interactive session whereby details about specific events are requested; a “cued recall” (e.g. “What can you tell me about the mouth?”).

To be similar with real life situations, after the exposure of a target, a CI based approach was used to elicit a description of the face. This involved two sessions of free recall followed by one session of cued recall (where participants are asked about each feature in turn). To maintain parallels further, the identity of the targets was hidden from the two operators and participants were requested not to reveal it.

Participants

Thirteen males and 17 females each created an EvoFIT. Their ages ranged from 15 to 55 and their mean age was 28.1 (SD = 9.3). They were paid £10. Eighteen males and 12 females each created an E-FIT. Their ages ranged from 18 to 51 and their mean age was 28.9 (SD = 9.5). Participation was voluntary.

Procedure

The basic procedure was kept the same for the creation of E-FIT and EvoFIT composites. Participants were told that they would be creating a composite of a famous face from memory. An envelope was given containing the targets and participants were instructed to remove one at random. If the person depicted was not familiar, they were told to replace the photograph and select another. When a familiar face was found, 1 minute was permitted for a detailed inspection of the target. The participants were asked not to reveal the identity of the famous person at any time during the session. The code (on the back of the photograph) was recorded and the target face placed in a second envelope that contained “used” stimuli.

A short description of the composite system was provided and an opportunity given for questions. Afterwards, a verbal description of the famous face was elicited, comprising of two cycles of free-recall followed by cued recall; details were noted on an E-FIT description sheet. A composite was then created using either E-FIT or EvoFIT. As normal, the operator controlled the composite software under the guidance of the witness to provide “a pictorial record of a witness’s memory and not that of the police artist or facial imaging operator” [ACPO(S) 2000], page 11). Finally, a percentage likeness to the target was estimated by the participant. Examples of composites constructed from each system may be found in Figure 6.

(a) E-FIT

(b) EvoFIT

Fig. 6. Examples of successfully identified composites from (a) E-FIT (Woody Allen, Michael Caine and Mick Jagger) and (b) EvoFIT (Bob Geldof, Nicholas Lyndhurst and Mick Jagger).

2.4. Evaluating the Composites

Evaluation primarily involved identification rates, achieved by asking another set of participants to recognize the celebrity composites. However, despite care taken to control for familiarity, concern was expressed that some celebrities may not be very familiar (e.g. the footballer, Michael Owen). Therefore, participants were also asked to name the original targets after they had finished naming the composites. This enabled a naming rate to be computed that was conditional on the number of targets known (i.e. a relative measure). As the important aspect of evaluation was recognition rather than naming, an unambiguous semantic description was acceptable. For example, “Big lips, oldie, lead singer in 60’s band,” would be taken as a correct response for Mick Jagger. This approach has been adopted elsewhere [Bruce et al. 1992]. The order of presentation was randomized for each participant.

Participants

Thirty-six participants volunteered to recognise the composites. These comprised of 12 males and 24 females. These were drawn from students attending the Open University summer school course D209, Stirling University plus staff and students at the University of Abertay.

EvoFIT

The composites were recognized a total of 39 times. As there were 540 presentations of the stimuli (18 participants * 30 composites), this resulted in a raw hit rate of 7.2%. If one divides the number of times a composite was recognized by the number of times the corresponding target photograph was recognized, a conditional hit rate (CHR) for each composite may be obtained. This procedure was adopted to compensate for differences in target familiarity (and to avoid unrealistically low naming rates – a potential problem with Davies and Oldman [1999]).

Thirteen composites were recognized by at least one person and the conditional hit rate (CHR) ranged from 0% to over 50%; the best recognition occurred for composites of Nicholas Lyndhurst (40%) and Mick Jagger (53%). The average CHR was 9.6% ($SD = 13.8$) and the average CHR of composites that were recognized by at least one person was 22.1% ($SD = 12.6$).

Figure 7 shows the conditional hit rate divided into the 3 distinctiveness categories. It can clearly be seen that the medium distinctness composites performed worse ($M = 5.7\%$, $SD = 13.1$) than both low ($M = 10.2\%$, $SD = 11.9$) and the high ($M = 13.0\%$, $SD = 16.4$) distinctive composites; high distinctive composites were recognized best overall. Inferential statistics for these data will be conducted in comparison with the E-FIT data later.

There was no significant correlation between the conditional hit rate and either the percentage likeness recorded at the end of the composite session ($r = 0.26$; $F(29) = 2.07$, $p = 0.161$), or the number of generations required to construct a composite ($r = -0.20$; $F(29) = 1.11$, $p = 0.301$).

Qualitative feedback from participants was positive regarding the Feature Shift Tool, the Facial Composite Tool and the shape and texture palettes.

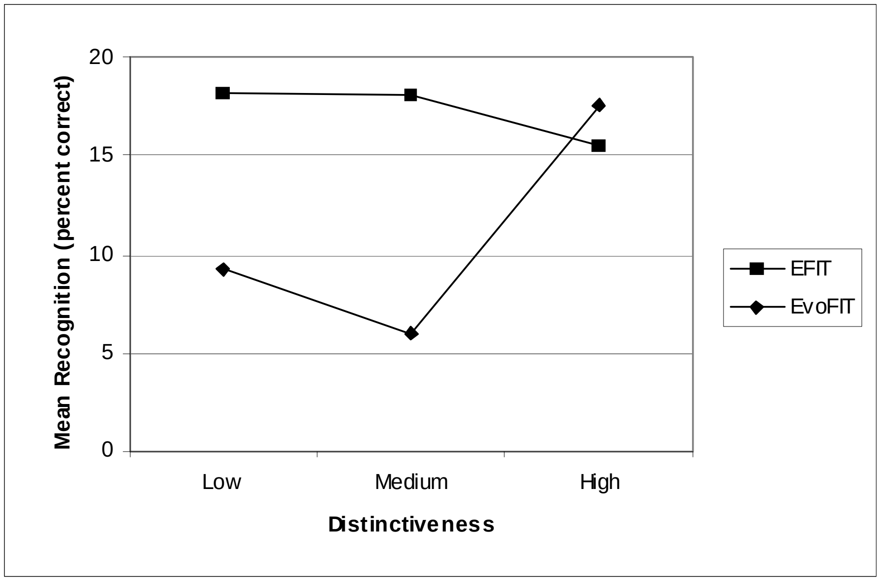

Fig. 7: Overall recognition (percent correct) between systems by distinctiveness for Experiment 1.

E-FIT

The E-FIT composites were recognized a total of 88 times. As there were 540 presentations of the stimuli (18 participants * 30 composites), this resulted in a raw hit rate of 16.3%. Twenty-two composites were recognized by at least one person and the CHR ranged from 0 to over 60%; the best recognition occurred for Woody Allen (61.1%). The average CHR was 17.1% (SD = 17.7) and the average CHR of composites that were recognized by at least one person was 22.6% (SD = 16.8).

Referring back to Figure 7, it can clearly be seen that the high distinctness composites (M = 15.6%, SD = 16.4) performed the worst and there was little difference between the low (M = 18.2%, SD = 19.5) or medium (M = 18.2%, SD = 18.9) distinctiveness categories. Once again, inferential statistics for the E-FIT data will be conducted in conjunction with EvoFIT in the following section.

As before, there was a low, non-significant correlation of 0.14 between the CHR and the percentage likeness recorded at the end of the composite session ($F(29) = 0.55$, $p = 0.465$).

Comparing E-FIT and EvoFIT

A two-factor mixed ANOVA [an analysis by-subjects with distinctiveness as a within-subjects factor and system as a between-subjects factor] indicated that there was a significant difference in the average CHR between E-FIT and EvoFIT ($F(1,34) = 9.43$, $p = 0.004$), there was no significant effect of distinctiveness ($F(2,68) = 1.42$, $p = 0.249$) but there was a significant interaction ($F(2,68) = 3.63$, $p = 0.032$). A simple-main effects analysis (of the interaction) revealed that E-FITs were named significantly better for the low distinctive ($p = 0.015$) and medium distinctive ($p = 0.002$) targets, but there was no significant difference between the systems for high distinctive targets ($p = 0.642$); there was also a significant increase in naming between medium and high distinctive targets for EvoFIT ($p = 0.024$).

2.5. Discussion

The analysis revealed that E-FIT composites were overall better recognised than EvoFIT composites but only for low and medium distinctive targets, with composites of high distinctive targets being equally well recognised. This suggests that were a witness to describe a highly distinctive face, similar performance would be expected whether E-FIT or EvoFIT were employed. The other interesting finding was that EvoFIT composites were recognised significantly better in the high distinctive category compared with the medium distinctive one. From this, one could argue that EvoFIT exhibited a distinctiveness effect, but, as performance was overall worse than E-FIT, perhaps a more appropriate interpretation is that EvoFIT appears to struggle with all but the most distinctive targets.

Interestingly, E-FIT performance is very similar to that found by Davies et al. [2000] under similar construction conditions (familiar targets constructed from memory). In the Davies study, a raw recognition rate of 17.3% was found, compared with 16.3% found here, indicating that naming rates of about 17% are likely with composites constructed from memory using E-FIT. We did not find a distinctiveness effect for E-FIT - unlike elsewhere in face perception [e.g. Shapiro and Penrod 1986] - and therefore it is perhaps the case that facial distinctiveness does not influence composite quality, a result that supports Green and Geiselman’s [1989] work with Identikit.

Target age is believed to be a factor causing lower overall recognition for EvoFIT (10%). Several participants did comment during the construction phase that the age of their EvoFIT appeared younger than that of the target. A small Internet-based study with 70 participants indicated that the average estimated age of the composites was 31.6 years ($SD = 8.8$). This was found to be significantly less than the mean age of the targets in this study ($M = 47.0$ years; $t(98) = 10.62$, $p < 0.001$). As such, this experiment does not represent a measure of likely system performance if used to create a composite of most suspects, who tend to be in their late teens and early twenties [Goffredson and Polakowski 1995].

3. EXPERIMENT 2

A second study was conducted to investigate whether EvoFIT was capable of creating age-appropriate composites (and exhibit an unambiguous distinctiveness effect). This was achieved by creating EvoFIT composites of young male Caucasian faces with the target-present. This was a deliberate decision to limit the effect of memory and focus evaluation on the capability of the system. Of course, it is appreciated that “in view” construction is likely to raise naming rates compared with constructions from memory but, importantly, the results are comparable to Brace et al. [2000] who constructed composites in a similar way with E-FIT.

3.1. Obtaining the Target Faces

Twenty good quality photographs of young famous faces (aged between 20 and 33 years) were obtained with occupations similar to the target set of Experiment 1. Twenty-eight participants were asked to rate how unusual they thought the faces were on a scale between 1 and 10. Participants were encouraged not to include familiarity judgements in their ratings. Five targets with low average distinctiveness (Craig Phillips, David Beckham, Noel Gallagher, Leonardo DiCaprio and Matt Damon) and five targets with high average distinctiveness (Robbie Williams, Michael Owen, David Schwimmer, Stephen Gately and Tim Henman) were selected. The mean age of these targets was 27.2 years (SD = 4.1) and the set was recognised on average 90% of the time. The mean distinctiveness rating was 4.6 (SD = 1.9) in the low distinctiveness condition and 6.6 (SD = 1.9) in the high distinctiveness condition; a significant difference ($t(27) = 8.80$, $p < 0.001$).

3.2. Creating the Composites

There were 5 operators, working mainly in small groups to create the composites: all 5 operators created Matt Damon; 3 operators created Michael Owen; one operator created Noel Gallagher and David Schwimmer; and the remaining composites were created by operators working in pairs. This was seen as a practical solution to aid in learning the EvoFIT software. Note that, contrary to intuition, no added benefit has been found for operators working in pairs [Davies et al. 1983] (note also that later work in this paper is experimentally better controlled using a single operator as part of a more realistic design, as in Experiment 1).

Composites were created for each of the 10 famous faces with a photograph of the target in view. Normal operating procedures were followed, starting with the initial selection of a hairstyle (PROfit) and then the selection and evolution of preferable faces. Composites took on average 5.9 generations (SD = 4.0) to complete.

3.3. Evaluating the Composites

Participants

Twenty-two participants volunteered, comprising of 10 males and 12 females. They were undergraduate students at the University of Stirling and had not taken part in the distinctiveness rating exercise.

Procedure

The EvoFIT composites were printed on separate A4 sheets using a high quality printer. As in Experiment 1, three tasks were given: (1) name the composites, (2) name the target photographs and (3) rate the quality of the composite in the presence of the target (a likeness rating); likeness rating was carried out using a 10 point scale (1 = Very poor likeness between faces, 10 = Faces are identical).

3.4. Results

All composites were recognized by at least one person and half of them were named over 20% (CHR); two examples may be seen in Figure 8. In total, there were 506 correct recognition attempts, resulting in an average hit rate of 25.3% (SD = 18.6). An analysis [by-subjects] revealed that the conditional hit rate for the high distinctiveness group (33.2%, SD = 22.5) was significantly greater than the low distinctiveness group (M = 17.4%, SD = 10.8, $t(21) = 3.72$, $p = 0.001$). Similarly, the average likeness ratings were significantly higher in the high distinctiveness condition (M = 4.6, SD = 1.5) than in the low distinctiveness condition (M = 3.8, SD = 1.3, $t(21) = 3.9$, $p < 0.001$). Due to the small number of items in each condition (5), by-items analyses were not conducted.

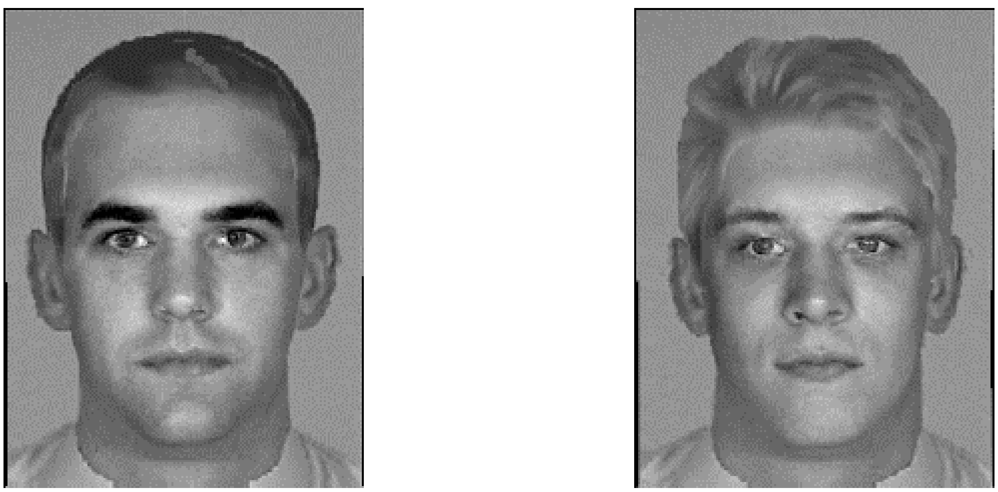

Fig. 8. Composites created using the EvoFIT system: the singer Robbie Williams (left) and the actor Matt Damon (right).

There was no significant correlation between the hit rate and either the number of generations taken to produce the composite ($r = -0.05$; $F(8) = 0.02$, $p = 0.892$), as found before, or the time taken ($r = -0.15$; $F(8) = 0.19$, $p = 0.678$). There was also a non-significant correlation between hit rate and the number of operators used to construct the composites ($r = 0.16$; $F(8) = 0.22$, $p = 0.649$), suggesting that the multiple operator approach does not diminish the validity of the results.

3.5. Discussion

This second study indicates that the EvoFIT system is capable of producing recognizable composites when the average age of the targets is more appropriate to the current database. All of the composites created were recognized by at least one person. Both the recognition rate and the likeness ratings illustrate a clear advantage when creating composites of distinctive faces using this system. We acknowledge that familiarity ratings of the target faces were not collected for this set, unlike the previous experiment, so perhaps composites of high distinctive targets were named better because they were more familiar to participants. However, we also found that composites of high distinctive targets were rated better for likeness, suggesting that familiarity was unlikely to have been (entirely) responsible for the “distinctiveness effect” in the naming data. Note also that the overall average likeness ratings were really quite low (4.3), fitting into the category of “Some similarities”. This does suggest that subjective quality can be low even though recognition can be quite good (25% overall).

4. EXPERIMENT 3

The next experiment involved performance with “witnesses” working from memory, as in Experiment 1. The design used a set of young celebrity faces (mean age of 26 years) but was improved to more closely reflect composite construction in real life. Firstly, participants studied a famous face that was unfamiliar to them (witnesses who construct composites do not know their assailant). Secondly, after seeing a target face, participants were required to wait 2 days before making the composite (a typical time interval for ‘real’ witnesses). Thirdly, to address possible differences in motivation, all participants were paid the same amount (£10) to be a “witness” (in Experiment 1, those using EvoFIT were paid but those using E-FIT were not). Finally, the evaluation also included PROfit, a composite system also used in the UK. As mentioned earlier, PROfit is very similar to E-FIT: they both contain a large database of facial features for witnesses to assemble. A third operator worked PROfit (this person was experienced, like the E-FIT and EvoFIT operators). Note that this design also reduces differences between PROfit and EvoFIT, since hairstyles came from a common source (PROfit).

Thus, 30 university staff and students looked at a photograph of an unfamiliar celebrity for 1 minute. Two days later, they were given a Cognitive Interview and constructed a composite with E-FIT, PROfit or EvoFIT (as in Experiment 1). The 10 composites from each system were given to a further 26 undergraduates to name (as before, participants also named the target photographs). Surprisingly, in spite of the target photographs being well named (M = 88%), only 2 composites from PROfit and 2 composites from EvoFIT were correctly named; there were no correct names for E-FIT. The resulting naming rate (CHR) was 3.6% for EvoFIT, 1.3% for PROfit and 0% for E-FIT. Although levels were very low, composites from EvoFIT were named significantly more often than the combined composites from PROfit and E-FIT ($p < .05$).

Clearly, the naming found here is strikingly less than composites constructed in Experiments 1 and 2. It would appear that a longer interval to construction and/or a younger aged target set produces very poor composites indeed. However, in spite of low composite naming, the results favored EvoFIT and suggest that it was at least as good as the other systems in a realistic setting.

5. EXPERIMENT 4

Even with more age appropriate targets, why were composites from EvoFIT not named better? We believe that the face model was a limiting factor. Firstly, there is clearly a problem with representation of eyes - the irises tend to be distorted away from their natural circular shape during the morphing process (notice the iris distortions in Figure 8 for Robbie Williams) - and such distortions may lead to construction and recognition deficiencies. A more accurate model has now been built, resulting in considerably better images (refer to Figure 9).

We also improved the initial generation of random faces. Recall that faces are generated by a weighted combination of eigenfaces. Weightings are Gaussian random numbers (mean of zero and unit standard deviation) scaled by the variance of the corresponding eigenface [computed from the PCA]. Occasional extreme values from the Gaussian distribution could result in unrealistic faces, wasting witness effort. A small trial demonstrated that a much better method is to scale a face coefficient to fit within the range of values used by the face model to represent the original face set (examples, Figure 9).

Fig. 9. Examples produced from the improved face model. Notice that distortions to the irises are reduced.

The current experiment served to explore these improvements. Once again performance was compared against PROfit using the realistic methodology of Experiment 3 (EvoFIT hairstyles were also obtained from PROfit). The design was improved by employing a single operator for both systems (more than one operator featured previously), a design that has been employed elsewhere to compare composite systems [e.g. Davies et al. 2000]. This has the benefit of avoiding differences between operators (such differences that are known to influence the quality of a composite [e.g. Davies, et al. 1983; Gibling and Bennett 1994]). Thus, the same person controlled both the EvoFIT and PROfit software.

For the construction stage, participants looked for 1 minute at a photograph of a footballer, unfamiliar to them but well-known to fans. Two days later, they described his face via a Cognitive Interview and constructed a composite using EvoFIT or PROfit (all ‘witnesses’ were paid equally). The resulting composites were given to 12 football fans to name. Once again, the overall level of naming (M = 6.1%) was low, but was higher for EvoFIT (M = 8.5%) than

PROfit ($M = 3.7\%$); a significant difference ($p < 0.05$). This indicates that the latest version of software does produce better composites than PROfit (despite poor naming rates); a result that supports Experiment 3.

6. OPERATION MALLARD

The system was used in a criminal investigation in the UK recently as part of “Operation Mallard”. This case involves a series of sexual offences carried out in Southern England over the last 3 years by a Caucasian male believed to be in his late twenties (all have been linked by DNA evidence). Sadly, despite considerable effort (including a public appeal), a PROfit and artist sketches failed to result in a conviction. Arguably, one problem concerned the likeness of the hair in the sketch for the third victim. This was due to the victim being unable to mentally form a clear image of the hair. Interestingly, a year later, after having seen a similar hairstyle on TV and then in the street, a clearer image could be formed. An updated sketch was then created for the hair alone and is now believed to be considerably better than the original. It was decided that an updated sketch of the hair be used as a basis for constructing an EvoFIT. Consequently, this sketch was resized, cropped and imported into EvoFIT.

An EvoFIT was constructed as normal by the selection of shapes and textures using this hairstyle. The witness required 3 generations to produce what she considered a very good likeness (Figure 10). During the composite session, the Facial Composite Tool – combining a shape from one face and a texture from another - was used twice. On both occasions, a preferable likeness was achieved. The first time, the resulting “composite” was selected as the best face for that generation. In addition, the Feature Shift Tool was used once: to reduce the inter-ocular distance by 4 pixels and to close the eyes by 2 pixels.

Evidence for the quality of the EvoFIT comes from (1) the high degree of satisfaction of the likeness expressed by the victim (and the considerable emotional response evoked) both after the composite session and 2 weeks later, and (2) the victim’s belief that the quality of the EvoFIT is superior to the original artist’s sketch and the PROfit. The EvoFIT is now being used as part of the police investigation and has been used in a public appeal (e.g. the BBC’s Crimewatch program).

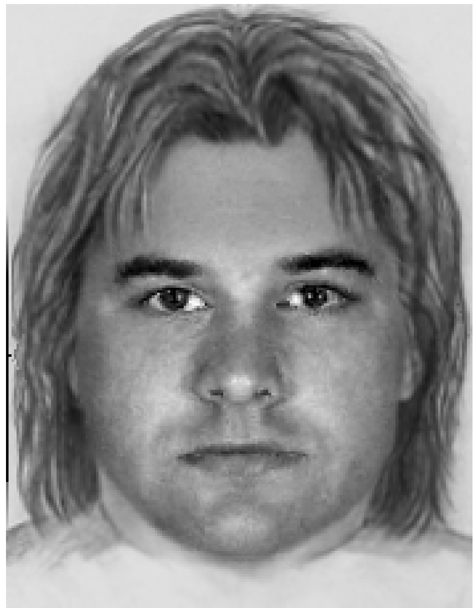

Fig. 10. The EvoFIT constructed in the field test. The hair, neck and shoulders were hand drawn by a Sketch Artist and imported into EvoFIT.

7. GENERAL DISCUSSION

Experiment 1 revealed that E-FIT composites were named better than EvoFIT composites. To account for the lower proportion of composites recognized, it was hypothesized that target age was a limiting factor. A follow-up study (Experiment 2) created EvoFIT composites of 10 targets with an average age of 27 years. This resulted in a hit rate of about 25%, the same rate found elsewhere for E-FIT with the target visible during construction. The next 2 experiments investigated EvoFIT using age-appropriate targets and a more realistic design (‘mock’ witnesses were unfamiliar with their target face and worked from memory). Both experiments produced low naming rates overall, but composites from EvoFIT were correctly named more often.

One might ask why the naming rate was not in general higher (10% in Experiment 1, 25% in Experiment 2 and just above floor level in Experiments 3 and 4)? We now believe that in spite of the improvements to the face model in Experiment 4, further work is still necessary to assist with face selection. The main problem is that faces appear very similar to each other - because each face is given the same hairstyle and has either the same texture (on the shape screen) or the same shape (on the texture screen) - and results in a difficulty selecting faces. This can be overcome in two ways. Firstly, by employing algorithms to increase the variability of the faces, for example by generating too many faces and then removing those most similar to others [e.g. Brown et al. 1994; Le Cun et al. 1990]. A second approach is to increase the number, and perhaps more importantly, the variety of faces in the face model. Currently, 72 faces are used, less than other PCA models [e.g. Sirovich and Kirby 1987; Troje and Vetter 1996]. Our aim is to augment the face model (to roughly double the number of faces used) and thereby increase the natural expressivity of the model. Steps are being taken to correct these issues and further realistic studies are planned.

Another area of fruitful work concerns the finding of Experiments 1 and 2 that there was no evidence of a correlation between composite naming rate and the number of generations (or time taken) to create the composites. One interpretation is that most of the useful work is done relatively early in the evolution process. A possible strategy, therefore, is to run the system for just a few generations, perhaps 2, and then to start again with a fresh set of random faces. If this process is repeated 2 or 3 times, then the best faces from each run could be brought together for further mixing and selection. Aspects of a face that are missing on one run might be present in another, allowing a better end result.

It is also an interesting open question whether it would be better to combine the results of several runs in this way, or to present the best face from each run, i.e. 3 or 4 faces simultaneously, for recognition. Previous studies suggest that either approach may be beneficial. McNeil et al. [1987] made a "modal" composite from the highest selected facial features used in 32 Identikits. They found that modal composites were rated significantly higher than their constituent Identikits. On the other hand, Bruce et al. [2002] found that simultaneously presenting composites from 4 people resulted in a 16% increase in identification rate compared with recognition from a single composite. We are currently investigating the use of both multiple witnesses and multiple images from one witness using EvoFIT.

Recognition data from sequences of best faces (including the composite itself) in Experiment 1 has been gathered from a further 18 subjects (also Open University students). Overall, this presentation format revealed a small lift in recognition, from 10.1% to 11.5%, with an especially large increase (from 18% to 50%) for the "age appropriate" Michael Owen (refer to Figure 11). This last finding does point to a possible method of composite presentation not naturally available to the other systems.

Fig. 11. Sequence of "best faces" selected during the construction of the composite for Michael Owen. Each was chosen at the end of each generation. The final image in the sequence, the "composite", is on the far right.

8. CONCLUSION

In Experiment 1, composites were constructed from memory and the overall naming rate of EvoFIT was less than E-FIT (except for highly distinctive faces). For the next experiment, the targets were more appropriately selected to be within the age of the database (about 30 years), and EvoFIT produced results similar to those of E-FIT in another study [Brace et al. 2000]. Two realistic studies were then conducted where mock witnesses constructed a composite from the memory of an unfamiliar face seen 2 days previously. Both studies found a low naming rate overall, but indicated that EvoFIT was superior to PROfit and E-FIT. Data from a field test also suggests continuing promise for EvoFIT.

Acknowledgements

This work was funded by EPSRC grant GR/L88627/01. The evaluation data for Experiment 2 was prepared by Kristine Green, Kimberley Malkin, Sarah McLanaghan, Jason Nairn and Kerry Ross. Bob Nichols at the UK Home Office provided the CD face database used to build the face model.

Bibliography

ACPO(S). 2000. *National working practices in facial imaging*. Association of Chief Police Officers (Scotland) Working Group.

BENNETT, P. 2000. The use of multiple composites in suspect identification. *Proceedings of the 3rd UK national conference on cranio-facial identification*. Manchester, May 2000.

BRACE, N., PIKE, G., and KEMP, R. 2000. Investigating E-FIT using famous faces. In A. Czerederecka, T. Jaskiewicz-Obydzinska and J. Wojcikiewicz (Eds.). *Forensic Psychology and Law*. Krakow: Institute of Forensic Research Publishers.

BROWN, G.D.A., HULME, C., HYLAND, P.D., & MITCHELL, I.J. 1994. Cell suicide in the developing nervous systems: a functional neural network model. *Cognitive Brain Research*, 2, 71-75.

BRUCE, V., HANNA, E., DENCH, N., HEALEY, P., and BURTON, M. 1992. The importance of "mass" in line-drawings of faces. *Applied Cognitive Psychology*, 6, 619- 628.

BRUCE, V., HEALEY, P., BURTON, A.M., DOYLE, T., COOMBES, A., and LINNEY, A. 1991. Recognising facial surfaces. *Perception*, 20, 755-769.

BRUCE, V., NESS, H., HANCOCK, P.J.B, NEWMAN, C., and RARITY, J. 2002. Four heads are better than one. Combining face composites yields improvements in face likeness. *Journal of Applied Psychology*, 87, 5, 894-902.

BRUNELLI, R., and POGGIO, T. 1993. Face recognition: features versus templates. *IEEE Transactions on PAMI*, 15, 10, 1042-1052.

CALDWELL, C., and JOHNSTON, V.S. 1991. Tracking a Criminal Suspect through “Face-Space” with a Genetic Algorithm. *Proceedings of the Fourth International Conference on Genetic Algorithms*, 416-421. Morgan Kaufmann Publishers.

COOTES, T.F., WALKER, K.N., and TAYLOR, C.J. 2000. View-based active appearance models. *Proceedings of the International Conference on Face and Gesture Recognition*. 227-232.

CRAW, I., and CAMERON, P. 1991. Parameterising images for recognition and reconstruction. *Proceedings of the British Machine Vision Conference BMCV '91*, Turing Institute Press and Springer Verlag.

DAVIES, G., MILNE, A., and SHEPHERD, J. 1983. Searching for operator skills in face composite reproduction. *Journal of Police Science and Administration*, 11, 4, 405-9.

DAVIES, G.M., and OLDMAN, H. 1999. The impact of character attribution on composite production: a real world effect? *Current Psychology: Developmental, Learning, Personality and Social*. 18, 1, 128-139.

DAVIES, G.M., VAN DER WILLIK, P., and MORRISON, L.J. 2000. Facial composite production: a comparison of mechanical and computer-driven systems. *Journal of Applied Psychology*, 85, 1, 119-124.

FIC 1999. *Crime management division: facial imaging course*. Scottish Police College.

FROWD, C.D. 2001. *EvoFIT: a holistic, evolutionary facial imaging system*. Unpublished Ph.D. Thesis, University of Stirling, Stirling, UK.

GIBLING, F., and BENNETT, P. 1994. Artistic enhancement in the production of photofit likeness; an examination of its effectiveness in leading to suspect identification, *Psychology, Crime and Law*, 1, 93-100.

GREEN, D. L., and GEISELMAN, R. E. 1989. Building composite facial images: effects of feature saliency and delay of construction. *Journal of Applied Psychology*, 74, 714-721.

HANCOCK, P.J.B. 2000. Evolving faces from principal components. *Behavior Research Methods, Instruments and Computers*, 32-2, 327-333.

HANCOCK, P.J.B., BRUCE, V., and BURTON, A.M. 1997. Testing principal component representations for faces. In J.A. Bullinaria, D.W.Glasspool and G. Houghton (Eds.). *Proceedings of 4th Neural Computation and Psychology Workshop*, London: Springer-Verlag, 84-97.

HANCOCK, P.J.B., BRUCE, V., and BURTON, A.M. 1998. A comparison of two computer-based face recognition systems with human perceptions of faces. *Vision Research*. 38, 2277-2288.

HANCOCK P.J.B., BRUCE, V., and BURTON, A.M. 2000. Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, 4-9, 330-337.

HANCOCK, P.J.B., BURTON, A.M., and BRUCE, V. 1996. Face processing: human perception and principal components analysis. *Memory and Cognition*, 24, 26-40.

KOVERA, M.B., PENROD, S.D., PAPPAS, C., and THILL, D.L. 1997. Identification of computer generated facial composites. *Journal of Applied Psychology*, 82, 2, 235-246.

LE CUN, Y, DENKER, J.S., & SOLLA, S.A. 1990. Optimal brain damage. *Advances in Neural Information Systems*, 2, 598-605.

LIGHT, L.L, KAYRA-STUART, F., and HOLLANDER, S., 1979. Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory*, 5, 3, 212-228.

MCNEIL, J.E., WRAY, J.L., HIBLER, N.S., FOSTER, W.D., RHYNE, C.E., and THIBAULT, R. 1987. Hypnosis and Identi-kit: a study to determine the effect of using hypnosis in conjunction with the making of identikit composites. *Journal of Police Science and Administration*, 15, 63-67.

O'TOOLE, A.J., VETTER, T., VOLZ, H., and SALTER, E.M. 1997. Three-dimensional caricatures of human heads: distinctiveness and the perception of facial age. *Perception*, 25, 719-732.

RAKOVER, S.S., and CAHLON, B. 1996. To catch a thief with a recognition test: the model and some empirical results. *Cognitive Psychology*, 21, 423-468.

SHAPIRO, P. N., and PENROD, S.D. 1986. Meta-analysis of facial identification rates. *Psychological Bulletin*, 100, 139-156.

SIROVICH, L., and KIRBY, M. 1987. Low-dimensional procedure for the characterization of human faces. *J. Opt. Soc. Amer. A*, 4, 519-524.

TROJE, N.F., and VETTER, T. 1996. Representation of human faces. *Technical report*, Max-Planck-Institute, Tubingen, Germany.

VALENTINE, T., and ENDO, M. 1992. Towards an exemplar model of face processing: the effects of race and distinctiveness. *Quarterly Journal of Experimental Psychology*, 44A, 671-703.